Homework 8: COMPAS and Algorithmic Fairness

36-313, Fall 2021

Due at 6 pm on Thursday, 28 October 2021

Agenda: Practice with the idea of algorithmic fairness; working with the black-boxed results of somebody else's statistical model.

Reading: The lecture on algorithmic fairness from 21 October

Our data set this week comes from the analysis, performed by the news organization ProPublica, of the "COMPAS" risk prediction scores for Broward County, Florida[1]. ProPublica compiled a data set on everyone arrested in Broward County over a certain time span, for whom the police or the jails had calculated a COMPAS score, and follow-up information about whether they had been re-arrested. (The course homepage provides further reading about this controversy. You don't *have* to read them for this assignment, but they can't hurt.) Specifically, our data file, `compas_violence.csv`, tracks the following information (in order):

- The age of each arrestee;
- Their age, binned into categories;
- Their sex;
- Their race;
- Their COMPAS score[2] for risk of violence (1–10, 1 being low and 10 high);
- Their COMPAS score, binned into categories of "Low" risk (1–4), "Medium" (5–7) or "High" (8–10);
- Their COMPAS score, binned into categories of "Low" (1–4) and "Medium or High" (5–10);
- Whether they were charged with a felony[3] (F) or misdemeanor (M);
- Count of priors[4]
- Whether they had a subsequent conviction for violence within two years. This is called "recidivism".

Notation: In this problem set, Y is the recidivism variable, 1 if the arrestee was re-arrested for violence within 2 years, and 0 otherwise. $\hat{Y}$ is the prediction of Y. The "positive" class will be recidivism, $Y = 1$, so a false positive means $Y = 0$ but $\hat{Y} = 1$, and a false negative means $Y = 1$ but $\hat{Y} = 0$.

1. *Understanding*

    a. (5) In a few sentences, using your own words, describe the data set in a way which should be comprehensible to a non-statistician. (You may want to actually look at the data file first.)

    b. (5) In a few sentences, using your own words, explain why one would want to build a statistical model to predict the risk of violence from features like this.

2. *Features and race*

    a. (5) Using histograms or other suitable graphics, show the distribution of (i) age, (ii) number of priors and (iii) COMPAS scores for (α) everyone, (β) blacks and (γ) whites. (You should have either a 3×3 array of plots, or 3 plots each with 3 curves.)

[1] Mostly: Fort Lauderdale, in the greater Miami metropolitan area.

[2] COMPAS calculates separate scores for risk of "failure to appear" at trial, risk of committing any type of crime, and risk of violence. We are only using the score for violence in this assignment.

[3] American law distinguishes between two kinds of crimes. Felonies are more serious crimes, punishable by (in most states) a year or more of imprisonment, or, in some situations, death. Misdemeanors are punishable by shorter terms of imprisonment (typically in city or county jails rather than state or federal prisons) and/or fines. Most crimes of violence are felonies, but not all felonies are crimes of violence: fraud, drug dealing, and tax evasion, for instance, are all felonies.

[4] This appears to be the count of prior *convictions* for crimes (not just arrests).

b. (5) How easy would it be to infer whether an arrestee was white or black from their age? From their number of priors? From their COMPAS score? Explain in words, referring to the plots you draw in (a). (Calculations are not required but are fine.)

c. (5) Is predicting redicivism from age just a disguised way of predicting recidivism from race? What about predicting recidivism from the number of priors? From the COMPAS score? Explain, by referring to parts (a) and (b).

3. *Accuracy and Error Rates of COMPAS* Suppose we predict recidivism for everyone whose COMPAS score reaches some threshold t, so $\hat{Y} = 1$ if $COMPAS \geq t$ and $\hat{Y} = 0$ otherwise. Since the scores are integers from 1 to 10, $t = 1$ would predict recidivism for everyone, and $t = 11$ would predict recidivism for no one.

   a. (5) *Accuracy* The **accuracy** of a statistical classifier is just the probability that it guesses the right class, $\mathbb{P}\left(Y = \hat{Y}\right)$. Plot the classification accuracy of the COMPAS score as a function of the threshold t. Include a horizontal line showing the baseline accuracy which we could achieve by predicting the same label for everyone (regardless of their score or any other features). For what thresholds (if any) does COMPAS improve on this baseline?

   b. (5) *FNR* The **false negative rate** of a classifier is $\mathbb{P}\left(\hat{Y} = 0 | Y = 1\right)$, in this case the probability that someone who *does* commit violence will be labeled non-violent. Plot the false negative rate of the COMPAS score as a function of the threshold t.

   c. (5) *FPR* Similarly the **false positive rate** of a classifier is $\mathbb{P}\left(\hat{Y} = 1 | Y = 0\right)$, the probability that someone who isn't violent will be labeled violent. Plot the false positive rate of the COMPAS score as a function of the threshold t.

   d. (5) *FNR vs. FPR* Plot the false negative rate against the false positive rate. (There should be 11 points on the plot, one for each value of t. [Or, if you make a line-type plot, the curve should have 11 corners.]) Describe the trade-off between the two types of error.

4. *Calibration of COMPAS*

   a. (5) For each level (1–10) of the COMPAS score, find the actual frequency of recidivism, i.e., what fraction of arrestees with that score were, in fact, violent recidivists. Do this separately for (i) everyone, (ii) blacks and (iii) whites. Plot the results. (One plot with three curves would be better than three plots.)

   b. (4) Repeat you plot from (a), but now add suitable error bars to all your estimated proportions. *Hints*: (i) If n trials each have success probability p, successes are independent across trials, and we observe x total successes, we can estimate $\hat{p} = x/n$, with approximate standard error $\sqrt{\hat{p}(1-\hat{p})/n}$. (What's "success" here? What's n?) (ii) `segments()` may be helpful for drawing.

   c. (5) Does the COMPAS score appear to be equally calibrated for both blacks and whites? Explain your answer by referring to the earlier parts of this problem.

5. *Fairness of COMPAS*

   a. (5) Predictions (or decisions more generally) are said to show **demographic parity** when the fraction of positive predictions is the same across groups. For races, this would mean that $\mathbb{P}\left(\hat{Y} = 1 | \text{Race}\right)$ is the same across races. Plot the fraction of arrestees with $\hat{Y} = 1$ as a function of threshold for (i) blacks alone, (ii) whites alone, and (iii) everyone. At what thresholds does COMPAS come closest to (or achieve) demographic parity?

   b. (5) Predictions have **parity of predictive accuracy** when they are equally accurate for different groups in the population. Re-do your plot of accuracy against threshold t, showing separate curves for whites, for blacks, and for the whole population. At what thresholds does COMPAS come closest to (or achieve) parity of predictive accuracy?

c. (5) Predictions have **parity of error rates** when error rates are equal across different groups in the population. Re-do your plot of false positive rates against threshold t, showing separate curves for whites, for blacks, and for the whole population. At what thresholds does COMPAS come closest to (or achieve) parity of false positives?

d. (5) Define the **violation of FPR parity** as the ratio of the false positive rate for blacks and the false positive rate for whites. Make a plot showing the violation of FPR parity against the accuracy. (This should have 11 different points [or corners], one for each value of t.) Describe the trade-off, if any, between parity and accuracy.

6. (10) *Advising Riverdale* Suppose that Riverdale County, Florida, is considering adopting COMPAS, and that you have been hired by a member of the county council to advise them about this decision. (You can assume that Riverdale County, while fictional, is otherwise very similar to Broward County, where the data come from.) Summarize what you have learned from this analysis about the ways in which COMPAS is or is not accurate and fair. Based on this, how would you recommend that the county use COMPAS, if at all? Would you recommend an alternative tool? Would it make a difference to your recommendation whether the council member was black, white, or something else?

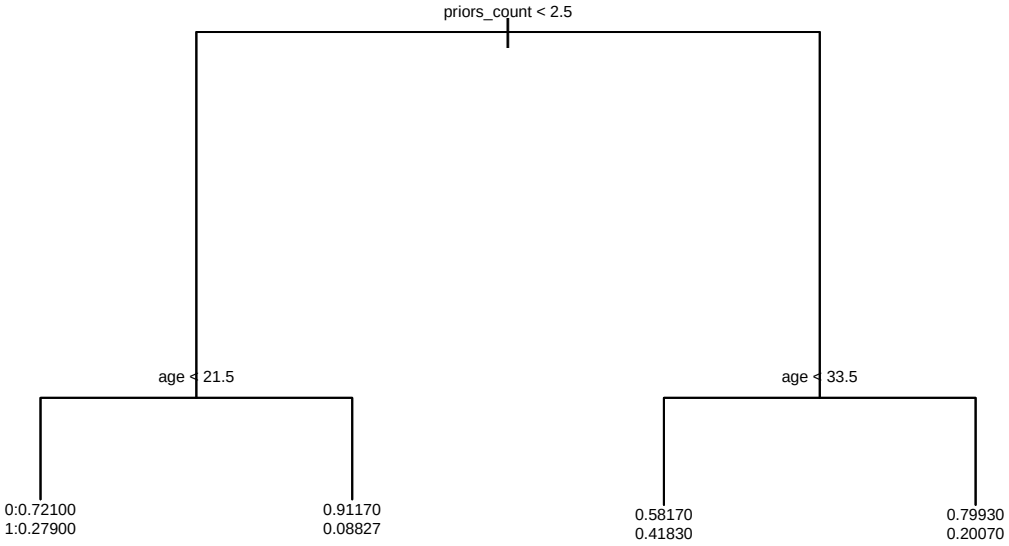
7. (1) *Timing* How long, roughly, did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Extra Credit

1. (5) We have assumed that if we use the COMPAS score, we need to apply the *same* threshold t to both whites and blacks. If we allowed there to be different thresholds for the two groups, could we achieve parity of false positive rates? If not, explain why not. If so, what would the common false positive rate be, what would the false negative rates be, and what would the accuracies be? Would you recommend doing this (assuming it's legal)?

2. (10) The question have asked you to look at whether COMPAS treats different races equally. We can also ask about whether it is fair across sexes. Re-do Q2, Q4 and Q5 to look at the disparity between the sexes rather than the races. Would this modify your conclusions in (6)? Why or why not?



3. The figure above shows a classification tree fit to this data set, with the goal of predicting recidivism. (This used the CART algorithm as implemented in the package `tree`.) The tree-growing algorithm had access to all of the variables in the data set (except the COMPAS scores), but didn't use all of them. It decided, in this case, to divide arrestees into four categories.
   a. (5) Describe, in words, the four categories, and the process which someone would step through to an arrestee to a category. What features of arrestees are used to assign them to categories? What is the predicted probability of violence for each category? Are any categories more likely violent than not? *Hint*: This part does not require you to fit the model yourself, just use the figure above.
   b. (5) Plot classification accuracy as a function of the threshold we apply to the predicted probability. Plot false negative and false positive rates. Is COMPAS any better at predicting violence than the classification tree? *Hint*: You could do this by re-fitting the classification tree and checking its predictions. But it's also enough to know that the four categories match (from left to right in the figure) 233, 2617, 557 and 613 cases.
   c. (5) Repeat Q5 for the classification tree. Is COMPAS any more fair, by those criteria, than the tree? If Riverdale has to use a risk-assessment algorithm, would you recommend paying for COMPAS, or using the tree for free?